APPLICATION NOTE

Comparison of Algorithms and Databases for Matching Unknown Mass Spectra

Fred W. McLafferty and Mei-Yi Zhang*
Chemistry Department, Cornell University, Ithaca, New York, USA

Douglas B. Stauffer and Stanton Y. Loh
Palisade Corporation, Newfield, New York, USA

The most used algorithms for the identification of electron-ionization mass spectra are INCOS and probability based matching (PBM). For unknown spectra of high purity, ~75% of rank 1 answers are correct for both algorithms, matched against the National Institute of Standards and Technology 62,235 spectrum database. With matching criteria that retrieve 50% of the possible correct answers from the Wiley 228,998 spectrum database, 54% of the PBM and 42% of the INCOS answers are correct; for 85% purity unknowns, 48% and 27% are correct. For an unknown spectrum of two compounds, neither was reported in the first three INCOS answers; eight of the first ten PBM answers identify both components. (J Am Soc Mass Spectrom 1998, 9, 92–95) © 1998 American Society for Mass Spectrometry

Electron ionization (EI) mass spectra are the most widely used data for unknown compound identification, with more than 10,000 gas chromatography/mass spectrometry (GC/MS) systems in worldwide use. Matching an unknown mass spectrum against a reference file [1] of 228,998 different spectra [275,000 including National Institute of Standards and Technology (NIST) spectra [2]], requires ~0.5 s [3], so that matching is a logical first step in identifying an unknown [4, 5]. These reference spectra were measured on a wide variety of instrumentation, so that the compound indicated should be verified by measuring its mass spectrum under the same experimental conditions as used for the unknown. For database matching, the most widely used computer programs are the Finnigan INCOS dot product [6] and the probability based matching (PBM) [7] algorithms.

A recent Stein and Scott study [8] of five mass spectrometry search systems concluded that the INCOS system (with their modified scaling, INCOS-SS) was the "best performing algorithm," achieving 75.7% accuracy for rank 1 answers versus 64.7% for their version of PBM (Table 1). This performance reexamination uses both the NIST [2, 9] and Wiley [1, 10] databases that contain multiple spectra (different sources) of the more common compounds, unknown spectra of varying purity, recall/reliability [11] as well as ranking evaluation, and the commercial version of the PBM algorithm [3].

Address correspondence to F. W. McLafferty, Baker Chemistry Laboratory, Cornell University, Ithaca, NY 14853-1301.

* Current address: Wyeth-Ayerst Research, Princeton, NJ.

Experimental

The INCOS algorithm used here is that described [6] plus that with the modifications of Stein and Scott [8], except for their undocumented prefilter that makes matching faster but less thorough. All components of the PBM algorithm [3] have been described extensively [4, 7]. Of reference spectra used, "exact duplicates" [4, 7] were removed from the 1992 NIST database [2] so that it contained 74,418 spectra of 62,235 different compounds. The *Wiley 1994 Registry of Mass Spectral Data*, 6th electronic edition [1], contains 228,998 different spectra of 198,348 different compounds.

The sets of simulated unknowns were the 12,593 alternate NIST spectra employed by Stein [8] and the 370 randomly selected spectra used in our previous studies [4, 7]. All have at least one other spectrum of the same compound in the data bases tested; a match of the unknown with itself is ignored. "Recall/reliability" evaluations [11] were made as described previously [4, 7]. For comparison with the Stein study [8], matches were counted as correct for reference spectra of the same Chemical Abstracts Service (CAS) Registry number, whereas "Classes I and IV" criteria [7b] were used for comparison with previous PBM studies. Class I counts stereoisomers as correct because they usually produce virtually identical mass spectra. Class IV matches are compounds differing structurally in ways that should only cause small variations in the mass spectrum. The ratio of Class I and Class IV matches found for earlier databases of 81,000 [7e] and 140,000 [4]

1044-0305/98/$19.00
PII S1044-0305(97)00235-3

Received June 5, 1997
Revised September 30, 1997
Accepted October 3, 1997

Table 1. Performance comparison based on highest rank answers

Database and algorithm	% correct at rank 1	1–2	1–3
62,235 database, 12,593 unknowns[a]:			
Ref. 8 data			
Commercial INCOS	72.9	85.9	90.8
INCOS-SS	75.7	88.0	92.5
PBM	64.7	78.4	84.8
This study:			
INCOS-SS	77.0	87.6	91.6
Commercial PBM[b]	74.9	83.2	86.4
228,998 database, 370 unknowns[c]:			
INCOS-SS	75.4	83.0	87.8
Commercial PBM	77.0	85.1	88.9

[a]Correct answers have same CAS number.
[b]If only the first reported answer is considered, even though a correct answer has the same RL value, the % correct are 72.8, 82.1, and 85.7.
[c]Class I correct answers.

spectra were very similar, so that this ratio was used here in calculating Class IV data.

Results and Discussion

Validity of Algorithm Implementations

Our version of the INCOS-SS algorithm was checked by repeating the Stein matching of 12,593 unknowns against the NIST 62,235 database. Their INCOS-SS ranking results [8, Table 2] are similar to ours (Table 1), with our slightly higher values consistent with the absence of the prefilter found to be "98% effective" [8]. In comparison, PBM uses "weighted file ordering" [7f] to achieve 0.2-s matching of this database.

With the commercial PBM algorithm [3], however, matching the 12,593 unknowns against the 62,235 database yields dramatically better results. Our rank 1 PBM answers are 74.9% (not 64.7% [8]) correct, comparable to 77.0% for INCOS-SS and 72.9% for the commercial INCOS [6] algorithms. Using the more comprehensive [11] recall/reliability (RC/RL) evaluation, PBM shows (Figure 1) a performance that is at least comparable with that of INCOS-SS; the minimum (adjustable) for PBM of RL > 0.1 causes its dropoff at >0.9 RC. By using the statistically representative unknown data set of 370 mass spectra selected at random [4, 8] and matching these against the full NIST 74,418 spectral database, again the INCOS and PBM algorithms give closely comparable performances (data now shown). The cause of these differences in their [8] and our PBM results is not obvious, but the commercial availability of the PBM algorithm should make verification straightforward.

For algorithm performance against the Wiley 228,998 database (Figure 2) using matching criteria in which 50% of the possible answers are retrieved (50% recall), 54% of PBM answers are correct versus 42% of INCOS.

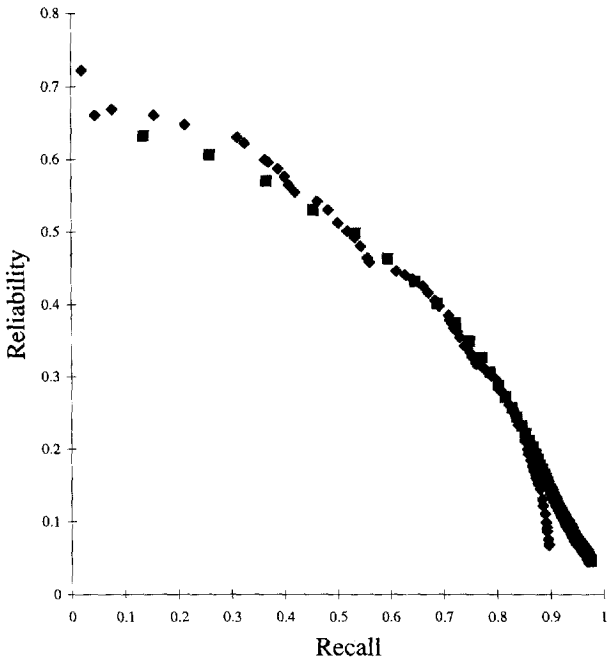

Figure 1. Recall/reliability results (same CAS number) of matching 12,593 NIST duplicate spectra against the NIST 62,235 spectra database using: square, INCOS-SS; diamond, PBM.

This better PBM discrimination against the far more numerous incorrect matches in the 228,998 database also results in a slight improvement in its "best answer" performance (Table 1).

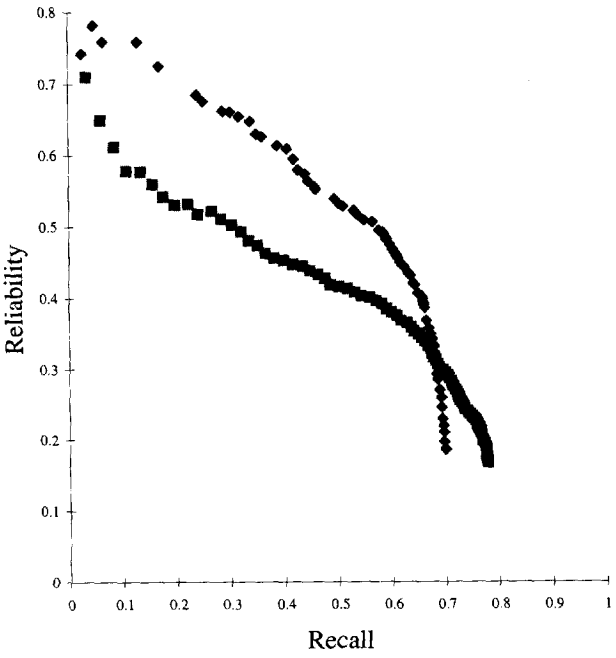

Figure 2. Matching (class I) 370 statistically selected unknowns against the Wiley 228,998 database using: square, INCOS-SS; diamond, PBM.

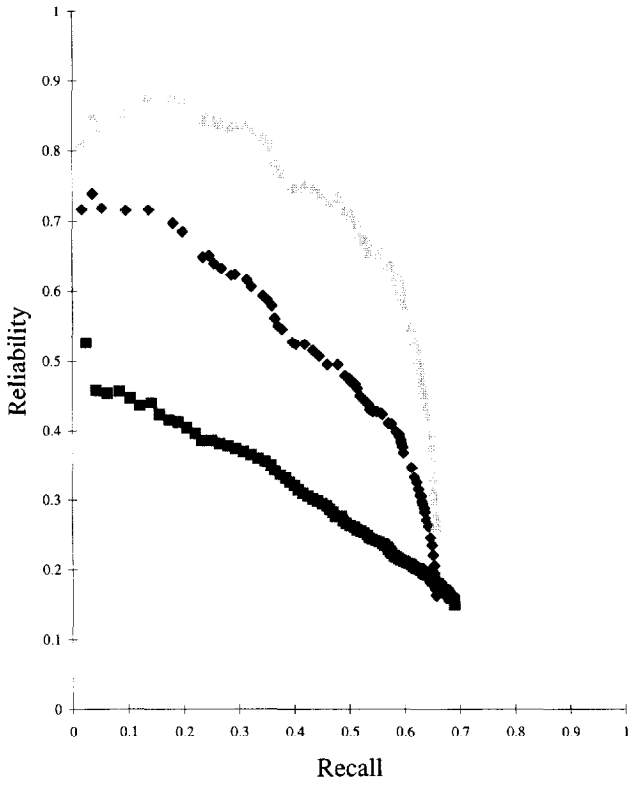

Figure 3. Matching the 85% component of 369 spectra synthesized as an 85:15 combination (similar molecular weights) of the 370 unknown spectra against the Wiley 228,998 database using: square, INCOS-SS, Class I; open diamond, PBM, Class I; triangle, PBM, Class IV.

Mixture Unknowns

Even the high separation capabilities of modern capillary gas chromatography can be compromised by mixture complexity, so that sample purity cannot be assumed. To test unknown spectra of lower purity, the 370 unknown mass spectra were combined by the computer to give 369 binary mixtures of similar molecular weight representing 85:15 proportions; these were matched against the Wiley 228,998 database (Figure 3). For PBM, this 15% impurity only reduces the RL value by 6% at 50% recall; 48% of the answers are correct (71% by Class IV criteria), while only 26% are correct by INCOS-SS (Figure 3). For 50:50 mixtures, a 25% recall yields 60% correct by PBM but 25% by INCOS. Matching the 85:15 mixtures against the NIST 74,418 database at 50% recall, 59% (Class I; 88% for Class IV) of PBM retrievals are correct versus 40% for INCOS-SS. Although the Wiley database contains an additional 154,580 possible wrong answers, the false retrieval of these is greatly reduced by the PBM reverse search [7] and spectrum subtraction [7d] capabilities.

Example Spectrum of an Unknown Mixture

Sparkman [12] recently reported an unknown spectrum (Figure 4) in which the multiple INCOS retrievals of

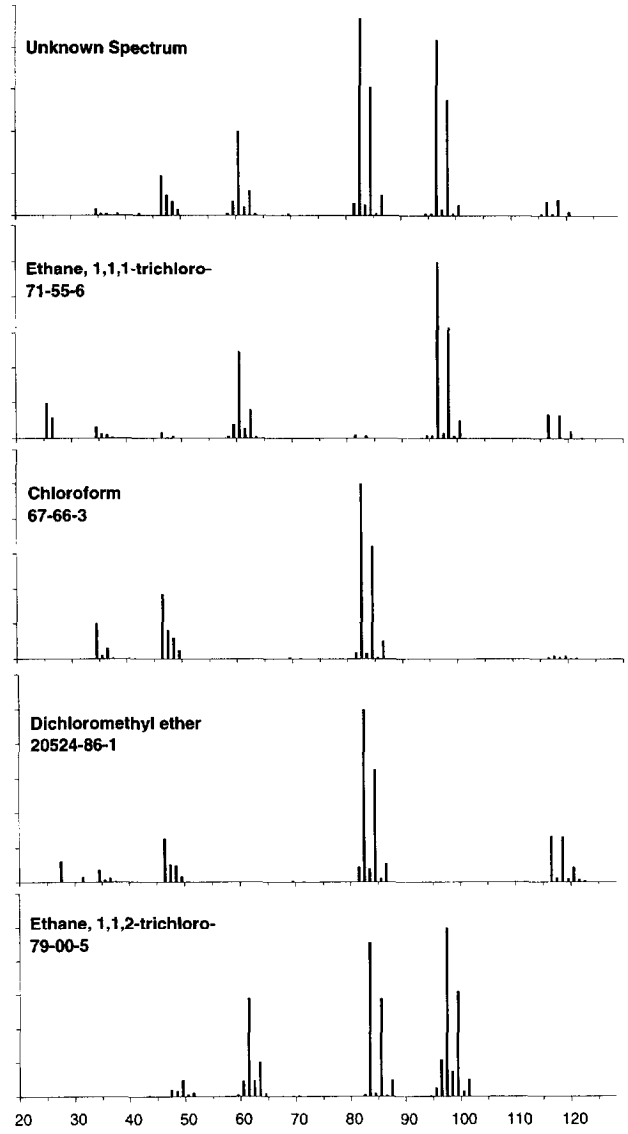

Figure 4. (Top) unknown spectrum of Sparkman [12] and (below) reference spectra (and CAS numbers) retrieved by PBM from the Wiley 228,998 database.

1,1,2-trichloroethane at high confidence levels from the NIST 74,817 database were all incorrect (Table 2). The unknown spectrum was from the top of a gas chromatographic (GC) peak; mass spectrometry spectra recorded near the base on each side of the GC peak showed that the unknown spectrum instead resulted from a mixture of 1,1,1-trichloroethane and chloroform. A PBM search (Table 2) does not retrieve 1,1,2-trichloroethane in the first 13 matches and indicates both of the correct components five times in its first ten hits. The first hit is incorrect, $Cl_2CHOCHCl_2$; its structurally misleading CCl_3^+ fragment ion apparently results from Cl and CH_2O losses. However, PBM with the Wiley 228,998 database (Table 2) gives a correct first hit, as well as correct answers representing both components for eight of the first ten hits with much higher reliability values than those from the smaller data base.

Table 2. Matching of Sparkman unknown mixture spectrum

Retrieved compound	INCOS: 74,827 [12] Rank: CL[a]	PBM: 74,418 Rank: RL[b]	PBM: 228,998 Rank: RL[b]
Correct:			
1,1,1-trichloroethane		2:61	1:79
		5:27	3, 4,5:64
		6:27	7:61
Chloroform		3:29	6:63
		10:17	
Incorrect:			
1,1,2-trichloroethane	1:838		8:56
	2:793		
	3:754		
Dichloromethyl ether		1:74[c]	2:78

[a]Confidence level: 800–900, high probability that unknown and reference are of same compound [6].
[b]Reliability value: % probability that reference is correct by Class IV criteria.
[c]Hits 4 and 9 are 1,1-dichloro-1-nitroethane, RL = 29 and 18.

Conclusions

These tests support the current routine use in many laboratories of the PBM algorithm [3] with the Wiley database of 228,998 spectra [1]. This requires ~0.5 s to match an unknown, and is especially advantageous for unknown spectra of impure samples.

Acknowledgments

David Sparkman provided the Figure 6 unknown and helpful suggestions; Caterina Stoenescu and Min-shuang Zou helped in the Registry database expansion. Previous support came from National Science Foundation grant no. CHE-8620293, and partial current support (FWM) came from National Institutes of Health grant no. GM16609.

References

1. McLafferty, F. W.; Stauffer, D. B. *Registry of Mass Spectral Data, 6th electronic ed.*; Wiley: New York, 1994.
2. NIST/EPA/NIH Mass Spectral Data Base, National Institute of Standards and Technology, Gaithersburg, MD, 1992.
3. Stauffer, D. B.; Loh, S. Y. *BenchTop Probability Based Matching*; Palisade Corporation: Newfield, NY, 1994.
4. McLafferty, F. W.; Loh, S. Y.; Stauffer, D. B. In *Computer Enhanced Analytical Spectroscopy*; Meuzelaar, H. C. L., Ed.; Plenum: New York, 1990; Vol. II, pp 163–181; McLafferty, F. W.; Stauffer, D. B.; Twiss-Brooks, A. B.; Loh, S. Y. *J. Am. Soc. Mass Spectrom.* **1991,** *2*, 432–437.
5. Warr, W. A. *Anal. Chem.* **1993,** *65*, 1045A–1050A, 1087A–1095A; Varmuza, K.; Werther, W. *J. Chem. Inf. Comput. Sci.* **1996,** *36*, 323–333.
6. Sokolow, S.; Karnofsky, J.; Gustafson, P. *The Finnigan Library Search Program: Finnigan Application Report 2*; Finnigan Corp.: San Jose, CA, March, 1978; Pellizarri, E. D.; Hartwell, T.; Crowder, J. A. *Comparative Evaluation of GC/MS Data Analysis Processing: Project Report PB-85-125664*; U. S. Environmental Protection Agency: Research Triangle Park, NC, 1985.
7. (a) McLafferty, F. W.; Hertel, R. H.; Villwock, R. D. *Org. Mass Spectrom.* **1974,** *9*, 690–702; (b) Pesyna, G. M.; McLafferty, F. W. In *Determination of Organic Structures by Physical Methods*; Nachod, F. C.; Zuckerman, J. J.; Randall, E. W., Eds.; Academic: New York, 1976; Vol. 6, pp 91–155; (c) Atwater, B. L., Stauffer, D. B.; McLafferty, F. W.; Peterson, D. W. *Anal. Chem.* **1985,** *57*, 899–903; (d) Stauffer, D. B.; McLafferty, R. W.; Ellis, R. D.; Peterson, D. W. *Anal. Chem.* **1985,** *57*, 771–773; (e) McLafferty, F. W.; Stauffer, D. B. *J. Chem. Inf. Comput. Sci.* **1985,** *25*, 245–252; (f) Mun, I. K.; Bartholomew, D. R.; Stauffer, D. B.; McLafferty, F. W. *Anal. Chem.* **1981,** *53*, 1938–1939.
8. Stein, S. E.; Scott, D. R. *J. Am. Soc. Mass Spectrom.* **1994,** *5*, 859–866.
9. Stein, S. E.; Ausloos, P.; Lias, S. G. *J. Am. Soc. Mass Spectrom.* **1991,** *2*, 441–443.
10. McLafferty, F. W.; Stauffer, D. B.; Loh, S. Y. *J. Am. Soc. Mass Spectrom.* **1991,** *2*, 438–440.
11. Salton, G.; McGill, M. J. *Introduction to Modern Information Retrieval*; McGraw-Hill: New York, 1983; Salton, G. *Science* **1991,** *253*, 974–980.
12. Sparkman, O. D. *J. Am. Soc. Mass Spectrom.* **1996,** *7*, 313–318.